


Interannual rainfall variability and SOM-based circulation classification

Piotr Wolski¹  · Christopher Jack¹ · Mark Tadross¹ · Lisa van Aardenne¹ · Christopher Lennard¹

Received: 21 October 2016 / Accepted: 4 March 2017 / Published online: 17 March 2017
© The Author(s) 2017. This article is an open access publication

Abstract Self-Organizing Maps (SOM) based classifications of synoptic circulation patterns are increasingly being used to interpret large-scale drivers of local climate variability, and as part of statistical downscaling methodologies. These applications rely on a basic premise of synoptic climatology, i.e. that local weather is conditioned by the large-scale circulation. While it is clear that this relationship holds in principle, the implications of its implementation through SOM-based classification, particularly at interannual and longer time scales, are not well recognized. Here we use a SOM to understand the interannual synoptic drivers of climate variability at two locations in the winter and summer rainfall regimes of South Africa. We quantify the portion of variance in seasonal rainfall totals that is explained by year to year differences in the synoptic circulation, as schematized by a SOM. We furthermore test how different spatial domain sizes and synoptic variables affect the ability of the SOM to capture the dominant synoptic drivers of interannual rainfall variability. Additionally, we identify systematic synoptic forcing that is not captured by the SOM classification. The results indicate that the frequency of synoptic states, as schematized by a relatively disaggregated SOM (7×9) of prognostic atmospheric variables, including specific humidity, air temperature and geostrophic winds, captures only 20–45% of interannual local rainfall variability, and that the residual variance contains a strong systematic component. Utilising a multivariate linear regression framework demonstrates that this residual variance can largely be explained using

synoptic variables over a particular location; even though they are used in the development of the SOM their influence, however, diminishes with the size of the SOM spatial domain. The influence of the SOM domain size, the choice of SOM atmospheric variables and grid-point explanatory variables on the levels of explained variance, is consistent with the general understanding of the dominant processes and atmospheric variables that affect rainfall variability at a particular location.

Keywords SOM · Synoptic circulation · Rainfall variability · Southern Africa

The relationship between synoptic circulation and local expressions of climate is the foundation of synoptic climatology and traditionally underpins the process of interpretative weather forecasting. The basic premise of synoptic climatology is that the large-scale atmospheric circulation exerts some control over weather and related environmental phenomena at the Earth's surface, thereby providing a level of deterministic forcing which influences the local weather. This directly leads to the use of atmospheric circulation patterns as a means to express drivers of weather (and climate) variability. Self-Organizing Maps (SOM, Kohonen 2001) is a pattern clustering method that is used as one of the methods to derive synoptic circulation types in this and other contexts, including statistical downscaling (e.g. Hewitson and Crane 2006; Yin et al. 2011; Ohba et al. 2016) and process-based validation of GCMs (e.g. Brown et al. 2010; Finnis et al. 2009; Higgins and Cassano 2010).

In the application of SOM to study climate variability, the evolution of synoptic drivers of weather is typically represented by progression through classes of daily synoptic states identified as SOM nodes. That progression

✉ Piotr Wolski
wolski@csag.uct.ac.za

¹ Climate System Analysis Group, University of Cape Town, Cape Town, South Africa

captures weather variability at daily time scales. In the top-down approach the mean weather (e.g. rainfall) is determined for the synoptic conditions expressed by each of the SOM nodes (e.g. Cassano and Cassano 2010; Engelbrecht et al. 2015; Engelbrecht and Landman 2016; Hewitson and Crane 2002; Schuenemann et al. 2008; Verdon-Kidd and Kiem 2009). In an alternative bottom-up approach, synoptic states corresponding to particular weather conditions, e.g. extreme or percentile rainfall, are identified (Cassano and Cassano 2010; Cavazos 2000; Lennard and Hegerl 2014). In quantitative terms, however, the day-to-day progression through synoptic states captured by SOM, explains a relatively low proportion of variance in weather responses. Cavazos and Hewitson (2005) show that for daily rainfall, the variance explained by synoptic states ranges between 20 and 80%, depending on location.

In the application of SOM to address variability at seasonal or interannual time scales, the day-to-day progression through SOM nodes translates into the year-to-year differences in SOM node frequency. Relatively few studies, however, have considered this aspect of SOM explicitly. Engelbrecht and Landman (2016) have shown that in South Africa's Cape region, anomalous frequencies of SOM-identified synoptic types mirror annual rainfall anomalies. Lennard and Hegerl (2014) revealed the correspondence between multi-year trends in the frequency of rain-bearing systems over South Africa identified through SOM, and rainfall associated with each of these systems. However, Hewitson and Crane (2002) noted that the trends in total precipitation and frequency of main rain bearing systems for a location in Pennsylvania did not correspond to each other, suggesting that the precipitation conditions and processes occurring within a circulation type, that are not captured by the SOM, may be important in determining interannual rainfall variability. and Hewitson (2005) suggested that these processes might be systemic, resulting in underestimation of wet, and overestimation of dry phases, and are perhaps related to the resolution of input datasets. These results are consistent with observations by Engelbrecht and Landman (2016) that the relationship between individual synoptic circulation types and rainfall is different during below-average and above-average years, with differences manifested through the intensity of rainfall events, and not just their frequency.

This paper uses SOM to identify sources of large-scale forced variability of rainfall at the interannual time scale, for locations in the winter and summer rainfall regimes over South Africa (Cape Town and Johannesburg). In doing so we address methodological questions stemming from Hewitson and Crane (2002) and Cavazos and Hewitson (2005). Firstly, to what extent the frequency of synoptic states captured by SOM explains seasonal and interannual variability in local responses? Secondly, is the residual

variability (i.e. not explained by the node frequency) in any way systematic, or is it an expression of in principle random, unresolvable processes? Lastly, is the rainfall-node frequency relationship sensitive to the design parameters of SOM, such as the choice of variables used to construct the SOM, the geographical domain and the dimensions of the SOM?

To address these questions, firstly, we quantify the relationship between the SOM-based frequency of synoptic states and local rainfall at the interannual time scale for locations in the winter and summer rainfall regions of South Africa—Cape Town and Johannesburg respectively. Subsequently, we use SOM to classify synoptic circulations for different domain sizes; for a typical large-scale (~35 deg) as well as for the smallest scale that may be considered meaningful in synoptic climatology (~6 deg). This allows us to test how domain size affects the ability of the SOM to disaggregate dominant synoptic drivers of rainfall over different rainfall regimes. We also use SOM classified using different sets of synoptic variables to understand how the choice of these affects the derived relationships with rainfall in each case. Additionally, we use a linear regression-based model to identify any residual variance (i.e. variance not already explained by changes in SOM frequencies) that is systematically attributable to variations in synoptic variables not already identified via the SOM. Together these results show that idiosyncrasies of the SOM method, combined with locally-dominant processes, are important to recognise when using SOM to express the synoptically driven variability (e.g. as part of a downscaling process), or evaluate the ability of climate models to simulate this variability.

1 Data and methods

1.1 Observed rainfall

Rainfall observations for stations falling within two “city-regions”, defined as the metropolitan areas of Cape Town and Johannesburg were obtained from the South African Weather Service (SAWS). Station data were quality controlled with tests for repeated, missing and unusual values. Stations with >95% data availability on a daily basis during the period 1979–2015 were selected for further analyses. This process left 27 and 12 stations for Cape Town and Johannesburg respectively. The analyses presented here were carried out for the dominant rainfall seasons in each case; (winter June–August) JJA for Cape Town, and (summer December–February) DJF for Johannesburg.

The general homogeneity of rainfall within each of the city-regions was confirmed by hierarchical clustering of time series of monthly rainfall totals; each city-region fell

within homogeneous zones which were different from other zones in the broader region. Despite this homogeneity daily rainfall differs between stations, partly due to the influence of local conditions (such as elevation or aspect) that locally modify the common synoptic forcing, and partly due to the chaotic nature of rainfall generating processes. Since our objective was to analyse the ability of SOM to capture the synoptic forcing of rainfall, we decided to work with region-average data rather than with individual station data. By using region average rainfall (and associated rainfall indices), the idiosyncratic behaviour of individual stations, as well as possible undetectable errors in observational data were averaged out, and only a regionally-consistent signal was retained.

1.2 SOM of circulation patterns

SOM, an artificial neural net-based method of topologically-sensitive clustering (Kohonen 2001), is often used in synoptic climatology for identifying large-scale synoptic circulation patterns (Hewitson and Crane 2002; Sheridan and Lee 2011). SOM is visualized as an array of data archetypes or nodes, which represents a nonlinear two dimensional mapping of circulation types. Distances between samples in the SOM “space” thereby represent degrees of dissimilarity in the original vector space (Hewitson and Crane 2002).

In this work, SOM of circulation pattern classes were based on daily 1979–2015 ERA-Interim (ERA-Int) reanalysis data (Dee et al. 2011). ERA-int data were subsetting for each of the domains, and subject to a principal component analysis (PCA) to reduce the dimensions of the data. The PCA was carried out separately for each variable. The number of retained components was determined based on randomization and assessment of significance using the N-rule test (Peres-Neto et al. 2005). Depending on variable and domain, between 2 and 9 components were retained. The retained eigenvectors were combined, and used as inputs to the SOM. SOM training used a 2-step approach (involving a coarse and fine search of the feature space, Hewitson and Crane 2002). It was determined that the procedure described above allows for obtaining SOM that have a high level of dissimilarity between different nodes—a preferred characteristic of SOM, visualized by a non-convoluted Sammon map (Sammon 1969).

Despite conducting the rainfall–synoptic forcing analyses only during the dominant rainfall season, SOM were trained on data for the entire year. The reason being that training on the entire year allows for differentiation of synoptic states that occur during transitional periods, i.e. towards the beginning and the end of the selected season, and are relatively rare, but might still be important rain-bearing states. In SOM trained on seasonal data only, states

in the ‘transitional’ period would likely be merged during training with other states.

A 7×9 SOM was used, the size of which was selected to represent a compromise between the level of required generalization and the number of available daily fields; a 63-node, 7×9 SOM allows that a single SOM node represents conditions occurring approximately 5 days per year (365 days/63 nodes), or 185 times in the training period of 37 years (i.e. 37 years \times 365 days/63 nodes).

SOM clustering was carried out on three sets of variables commonly used to define the synoptic state: (1) specific humidity (q), air temperature (t), zonal (u) and meridional (v) wind, (2) geopotential height (z) and (3) combined q , t , u , v and z . For simplicity, thereafter we refer to these three different types of SOM as the $qtuv$, z and $qtuvz$ SOM respectively. For Cape Town, the initial analyses were carried out using variables at 850 hPa, while for Johannesburg at 700 hPa. This is motivated by the difference in elevation of the regions: Cape Town is approximately at sea level, while the Johannesburg region is at 1400 m a.m.s.l. Thus the different pressure levels represent a similar height above ground. SOM on variables at the other levels were used in auxiliary analyses.

1.3 Domain selection

Two differently sized domains (large and small) were used for each location. The large domains were defined so that the dominant circulation systems influencing each of the two regions, as described below, are well captured. In southern Africa, the dominant influence on local weather, and particularly rainfall, comes from the interaction between the sub-tropical anticyclonic high pressure belt and perturbations in the temperate westerlies forming cyclonic low pressure systems. These hemispheric-scale systems migrate in a north–south direction between austral summer and winter and influence the weather in the two cities differently (Tyson and Preston-White 2000).

Over Cape Town, Austral summer (DJF) experiences dry, windy conditions as a result of the dominating influence of the more poleward location of the sub-tropical anticyclonic high pressure belt. During winter (JJA), the high pressure belt moves equatorward and mid-latitude cyclones with associated cold fronts, “sweep” across the tip of the continent, bringing rainfall to Cape Town and its region (Tyson and Preston-White 2000). Other significant rain-delivering synoptic systems over Cape Town include cut-off lows, but these are usually smaller in size (Favre et al. 2013), and thus less relevant to the selection of the domain. The large domain, which was chosen to capture both the latitudinal migration of the high-pressure systems, and longitudinal propagation of low pressure cyclones, extends

from 46° to 20°S and 0° to 30°E (Fig. 1), and is named here *wcp* (for Western Cape).

Rainfall in Johannesburg occurs predominantly in austral summer (DJF) and is delivered mostly in the form of convective thunderstorms. Its occurrence is related to the interaction between the Indian Ocean sub-tropical high pressure system, bringing in moisture to the subcontinent by tropical easterlies, and a weak continental thermal low (Tyson and Preston-White 2000). This basic mechanism can be enhanced through conditions facilitating development of moisture fronts, or so-called tropical temperate troughs (TTTs), where disturbances in the tropical easterlies interact with those in the subtropical westerlies creating linear NW–SE convective cloud band systems (Harrison 1984; Hart et al. 2013). The conditions facilitating TTTs commonly occur when the easterly tropical low over the interior connects with the westerly wave or a cut-off low to the south of the continent. The large domain that captures the relative position and shape of the high pressure anticyclone over the southern Indian ocean, and the low pressure systems to the south of the continent extends from 46° to 10°S and from 10° to 56°E (Fig. 1), and is named here *saf* (for southern Africa).

Small-scale domains named *cpt* for Cape Town and *jhb* for Johannesburg, were sized 6 by 6° and centered over the two cities (Fig. 1). The size of these domains were chosen to represent the ‘smallest’ possible representation of synoptic conditions that would capture gradients in the atmospheric circulation i.e. 3×3 grid boxes of the 2° ERA-Int reanalysis. A similar domain size is used, for example, in the statistical downscaling procedure of Hewitson and Crane (2006).

Despite less information (or variance) being contained in the small domain data compared to the large domain data, a SOM of identical size (i.e. 7 by 9 nodes) was used

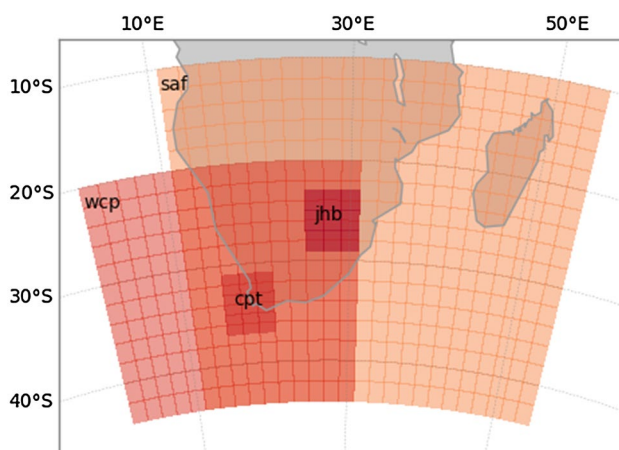


Fig. 1 SOM domains representing large (*saf*) and small domain (*jhb*) for Johannesburg, large (*wcp*) and small domain (*cpt*) for Cape Town

in both cases. This was dictated by the need to avoid the influence of the size of SOM on the relationship between SOM-schematized synoptic state and rainfall, thus enabling us to clearly identify the influence of the size of domain on the results.

SOM archetype maps were obtained through averaging the mean sea level pressure field on days when a given node occurred (Fig. 2). Despite the use of dimensionally-reduced (i.e. subject to PCA) datasets, training of SOM for the large domains generated solutions that were often not stable, i.e. different initializations of training with identical input data and training parameters produced slightly different SOM. As reported by other researchers (e.g. Sheridan and Lee 2011), the process of SOM training does not always converge on a unique solution. Practice of SOM implementation often relies on trial and error and the experience of researchers. Here, after the initial trial-and-error attempts at SOM training with our datasets, we settled on the approach outlined above as the one that best suited our data, but to account for the inherent uncertainty due to the training (identification of archetypes) we trained 20 SOM for each dataset/domain.

1.4 Analyses of relationship between SOM classes and interannual rainfall variability

We illustrate the factors underlying interannual variability in seasonal rainfall that are resolved and unresolved by SOM through mapping differences between the 10 wettest and 10 driest JJA (for Cape Town) or DJF (for Johannesburg) seasons in the 1979–2015 period in terms of:

- Frequency of nodes (number of days a given node occurred, or “node days”, per season)
- Total per-node rainfall (total rainfall recorded on node days per season)
- Mean daily rainfall on rain days (mean daily rainfall on node days with recorded rainfall)
- Fraction of rain days, per node (ratio of node-days with recorded rainfall, to node days)

Indices reflecting the above were calculated using data from individual stations, and their values averaged over each region. The wettest and driest seasons were selected based on total seasonal rainfall.

We construct a linear regression model (ordinary least square regression) for each node, relating total seasonal rainfall recorded during node days to node frequency, i.e.:

$$P_n = \alpha_n + \beta_n \cdot F_n$$

where P is the seasonal rainfall total, F is seasonal node frequency, α and β are intercept and regression coefficient respectively, and the superscript n denotes SOM node.

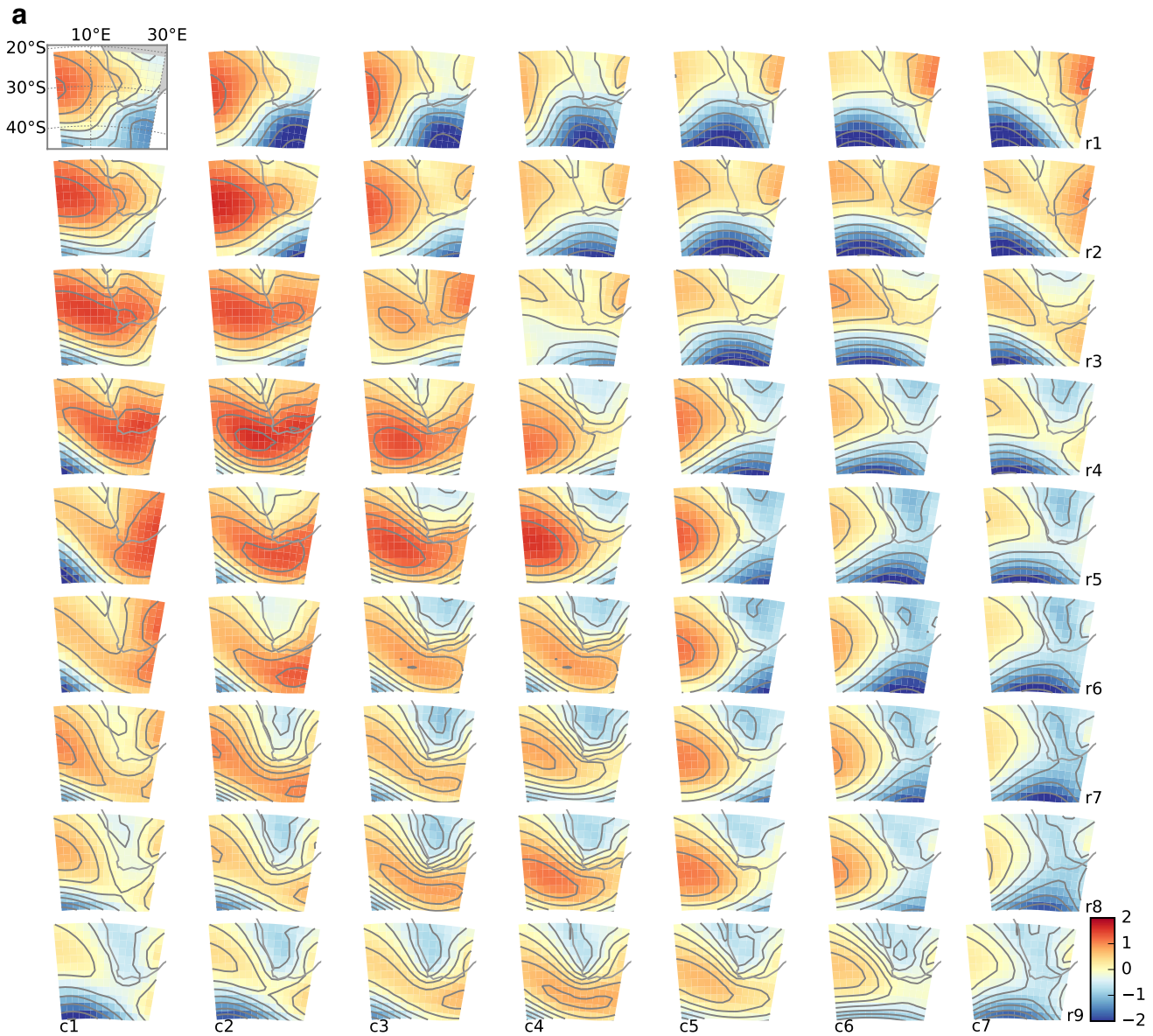


Fig. 2 Archetype maps of standardized anomaly of sea level pressure for 7×9 *qtuw* SOM for *awcp* domain and *bsaf* domain

This piecewise regression allows us to calculate the predicted total seasonal rainfall as a sum of individual node rainfall values:

$$P = \sum_{i=1}^N P_n$$

The square of the correlation coefficient between the predicted and the observed rainfall time series expresses the fraction of total variance that is explained by a given regression model. While analysing the variance of seasonal rainfall totals, we consider it to be composed of two elements: systematic and random. The systematic component is an expression of synoptic forcing that our regression

models attempt to capture, and will manifest through correlations between seasonal rainfall totals and synoptic variables, as well as through the systematic character of regression model residuals. The random component is variance that is unresolvable within the synoptic climatology framework, arising due to small-scale feedbacks and the random nature of atmospheric processes triggering rainfall events.

In order to elucidate synoptic variables that drive interannual variability and whose total variance is not captured by SOM frequencies alone, we use multiple linear regression in the form:

$$P_n = \alpha_n + \beta_n \cdot F_n + \gamma_n \cdot x_n$$

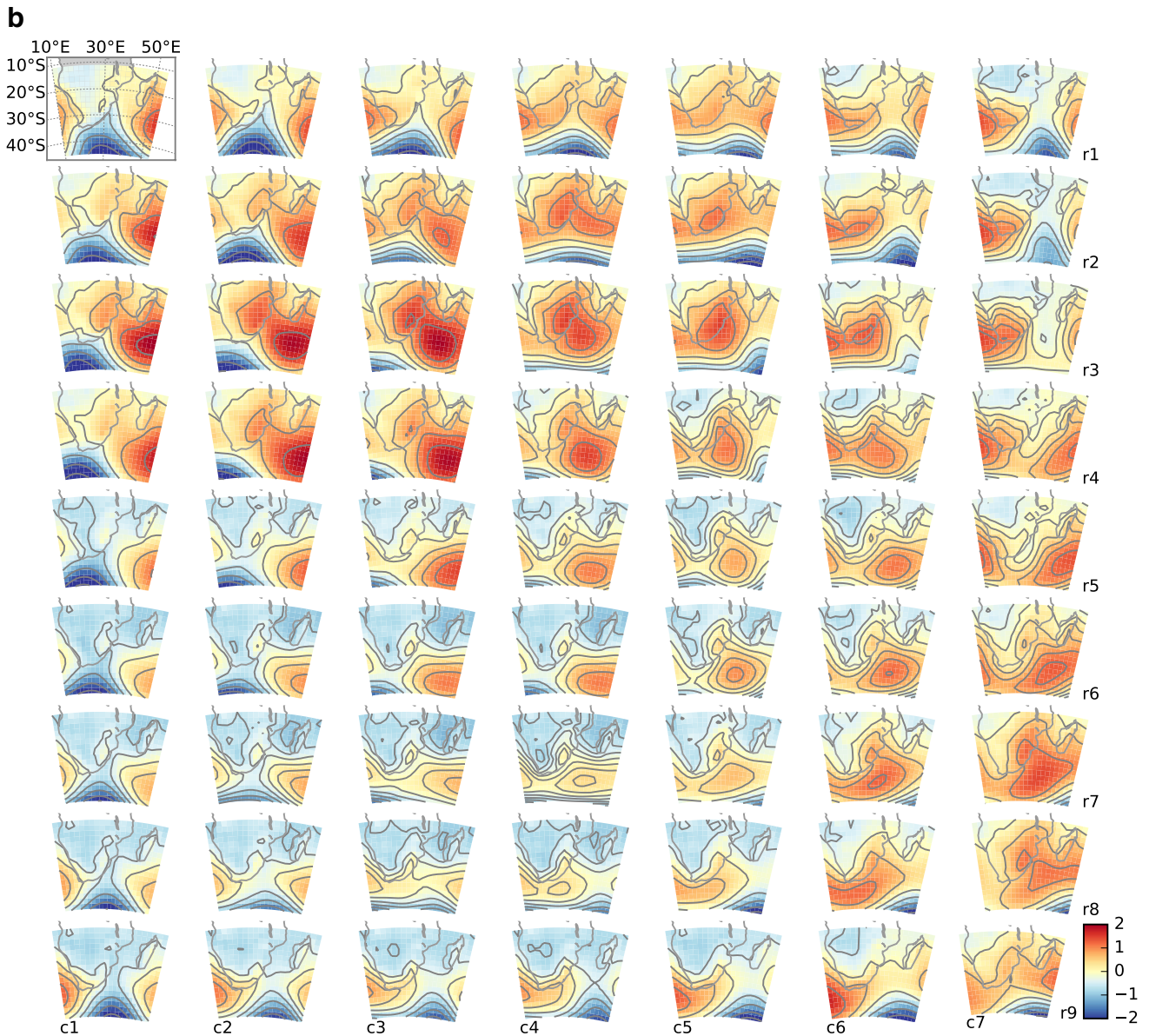


Fig. 2 (continued)

where x is an additional explanatory variable, and γ a regression coefficient associated with that variable.

We consider grid point values of a number of synoptic variables: q , t , u , v and horizontal moisture divergence (div) at 700 and 850 hPa levels, the latter calculated from u , v and q . Additional variables are the temperature lapse rate ($lapse$), calculated as the difference between air temperature at 700 and at 850 hPa, convective available potential energy ($CAPE$) obtained directly from the ERA-Int archive, and mean sea level pressure ($mssl$). In the multiple regression analyses, the values of these variables were derived for node days, and averaged to seasonal means for a particular node.

In order to investigate how the choice of SOM domain and choice of synoptic variables influence the relationship between SOM node frequency and seasonal rainfall, different sizes of the SOM domain and different synoptic variables were used in the training of SOM. These experiments were carried out in a similar configuration to those presented above, i.e. piecewise linear regression of seasonal rainfall totals against SOM node frequency, which was implemented for each combination of domain size and synoptic variables. A multiple linear regression analysis was then conducted that included each of the additional synoptic explanatory variables, as well as the SOM nodal frequencies. This procedure allowed for assessment of

differences in explanatory power of SOM node frequencies, as well as the gains in representing interannual variability provided by inclusion of a single additional synoptic variable in the regression.

2 Results

2.1 Circulation patterns associated with dry and wet conditions

The relationship between circulation patterns captured by the large domain *qtuv* SOM and rainfall is different for Cape Town and for Johannesburg (cf. Figs. 2, 3).

For Cape Town, the *qtuv* SOM for the *wcp* domain (Fig. 2a) clearly distinguishes between rain-bearing and primarily dry circulation types occurring during the JJA season (cf. Figs. 3a, 4a). The rain-bearing states are those associated with passing cold fronts, occupying the r1 and r2 (for reference to SOM nodes we use here row-column, or r-c, coordinates, with rows counted from top, and columns from the left hand side of the SOM grid) of the SOM in Fig. 2a, with r1 capturing the temperate low pressure system in its most northerly, and r2—in its more southerly position. The Cape region also receives some rainfall under conditions represented by node r7-c1 and its neighbours in Fig. 2a. These conditions likely represent a high

pressure system that typically follows a cut-off low pressure system translating north-eastwards across the country. The high advects moisture from the south and sometimes causes rainfall over the study region, albeit lower than during the passage of cold fronts.

The predominantly dry states occurring in JJA are those associated with troughs of a westerly wave in the temperate low pressure belt, and the sub-tropical anticyclone ridging into the Cape region. These states occupy r4-c1:c5 part of SOM in Fig. 2a. The remainder of the SOM nodes, i.e. the bottom-right region represents conditions where low pressures dominate over the continent, potentially associated with rainfall in the interior. These conditions occur typically during the DJF season.

For Johannesburg, the *qtuv* SOM for *saf* domain (Fig. 2b) captures a wider variety of wet states during the DJF season, or in other words, there is less “sharpness” in the circulation–rainfall relationship than over the *wcp* region (cf. Figs. 3b, 4b). The circulation depicted by the most frequent DJF SOM nodes in Fig. 2b (r5:r9-c1:c5) is clearly characterized by the low pressure system over the subcontinent, as well as over the tropical west Indian ocean, with the individual nodes capturing differences in the position, depth and the general shape of the Indian Ocean sub-tropical high pressure system, and the general progression of both the temperate westerly and tropical easterly waves. Although these circulation features are seemingly the

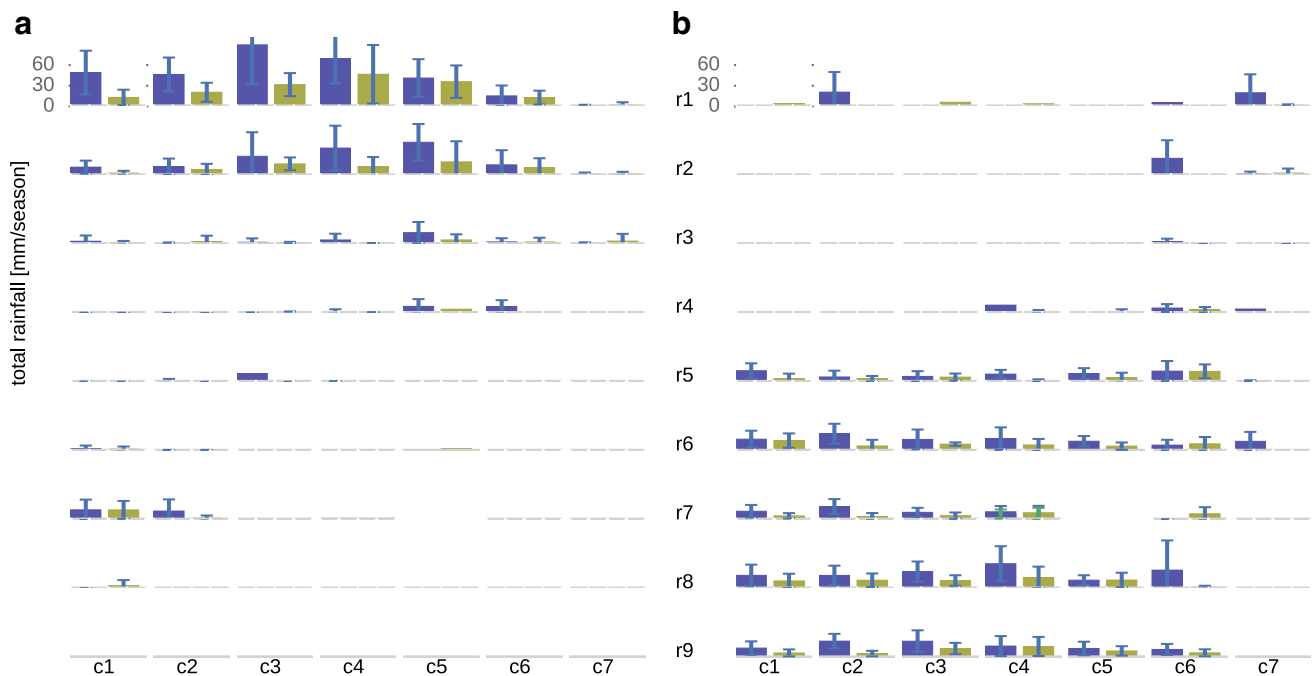


Fig. 3 Total rainfall occurring during days characterized by synoptic conditions schematized by SOM shown in Fig. 2, for **a** Cape Town (during JJA) and **b** Johannesburg (during DJF). Bars are mean value

in 10 wettest (in blue) and 10 driest (in orange) JJA or DJF seasons in the 1979–2015 period, error bars denote ± 1 standard deviation range

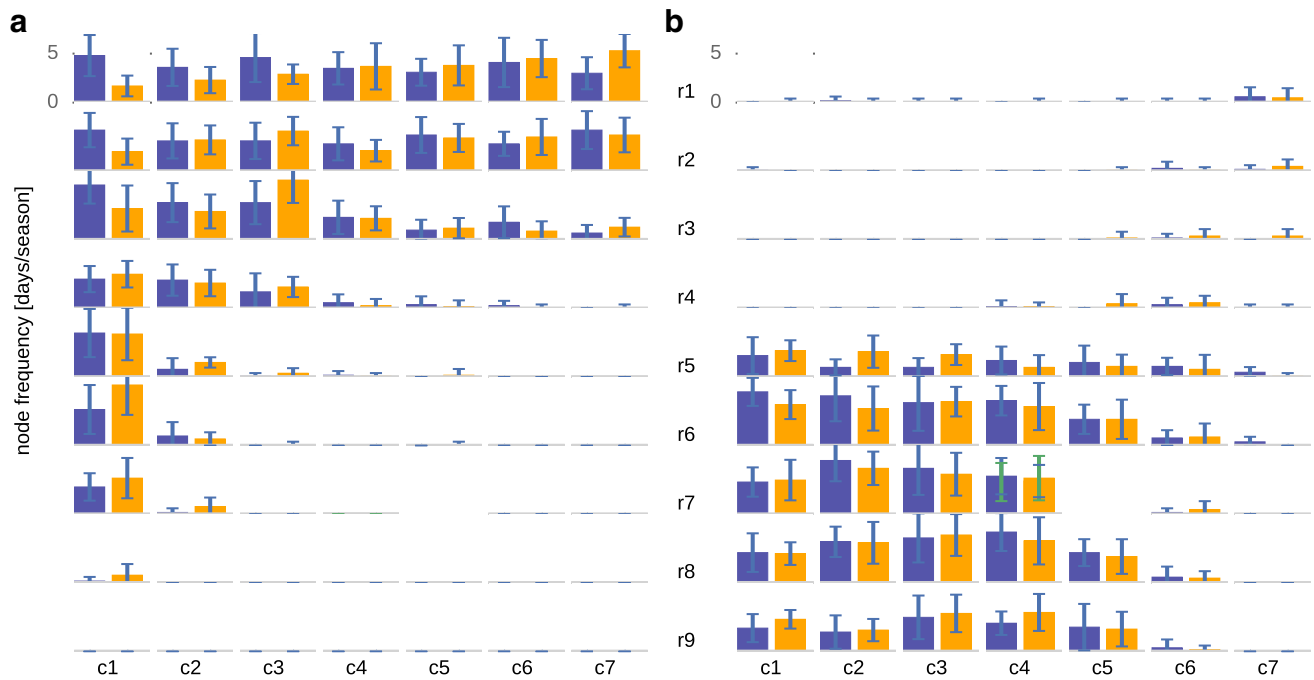


Fig. 4 Frequencies of synoptic states schematized by SOM shown in Fig. 2, for **a** Cape Town (during JJA); and **b** Johannesburg (during DJF). Bars are mean value in 10 wettest (in blue) and 10 driest

(in orange) JJA or DJF seasons in the 1979–2015 period. Error bars denote ± 1 standard deviation range

driving forces of rainfall variability in the Johannesburg region, the SOM training struggles to differentiate between rain and no-rain conditions. This can be seen in that rain events of similar magnitudes are observed under all synoptic states occurring frequently during the DJF season (cf. Figs. 3b, 4b), and there are no “dry” DJF states. This is perhaps not surprising given that sub-grid scale convective systems are responsible for rainfall over Johannesburg. Besides the circulation states characterized by the predominantly low pressure system over the sub-continent’s land mass, Fig. 2b also depicts several states that occur less frequently during the DJF season. These are characterized by a strong subtropical Indian Ocean high, as well as a relatively high pressure system over the tropical Indian Ocean (c6:c7), but a relatively low pressure system over the western sub-continent. These states, although infrequent, may facilitate high intensity rainfall events over Johannesburg (Fig. 5b) by bringing moisture directly from the Indian Ocean region over Johannesburg, and as a result be responsible for a relatively high proportion of the total rainfall. Nodes in c1-r5:r8 and c2-r6 (Fig. 2b) show a trough connecting the mid-latitude with the sub-tropical low pressure system, suggesting synoptic states that may facilitate the formation of tropical temperate troughs (TTTs).

2.2 Differences between dry and wet years

In both large domain SOM there are clear differences between the per-node total amount of rainfall delivered in wet and dry seasons (Fig. 3), though often these differences are small compared to the standard deviation between years i.e. these differences are not necessarily statistically significant. For Cape Town, the wet–dry differences in per-node total rainfall are matched by the differences in nodal frequency (see Fig. 4a), i.e. wetter years tend to have more frequent occurrence of nodes with high rainfall totals (e.g. r1-c1:c3), while dry years tend to have a higher frequency of nodes characterized by low or no rainfall (e.g. r1-c5:c7). For Johannesburg such a relationship is not clearly evident (Fig. 4b). Over both domains, however, wet years are characterized by higher rainfall intensities during days with rain (Fig. 5) and a slightly higher frequency of rain days (Fig. 6) than the dry years under identical circulation conditions. With a few exceptions (e.g. node r1c7 in Fig. 5b), these relationships are weaker for Johannesburg than for Cape Town. The wet–dry year differences in per-node rainfall intensity and rain day frequency clearly indicate the presence of a systematic rainfall forcing that is not captured by the SOM circulation patterns alone.

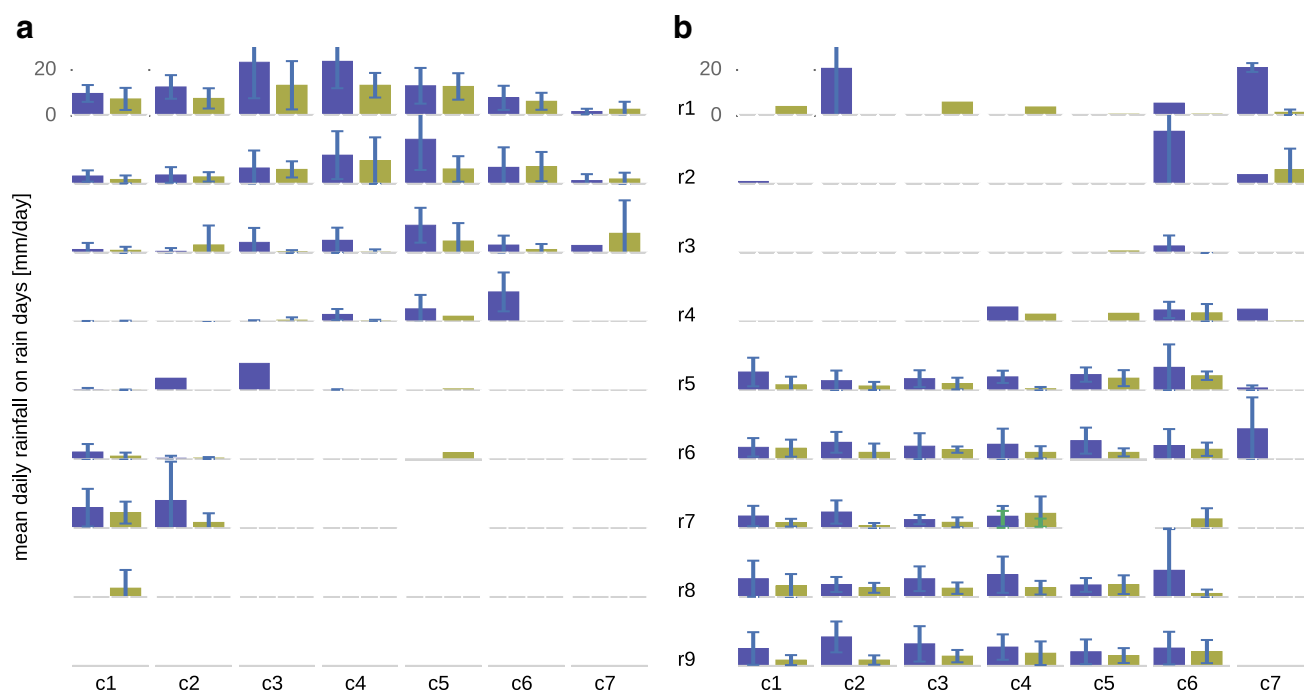


Fig. 5 Mean daily rainfall on rain days characterized by synoptic conditions schematized by SOM shown in Fig. 2, for **a** Cape Town (during JJA) and **b** Johannesburg (during DJF). Bars are mean value

in 10 wettest (in blue) and 10 driest (in orange) JJA or DJF seasons in the 1979–2015 period, error bars denote ± 1 standard deviation range

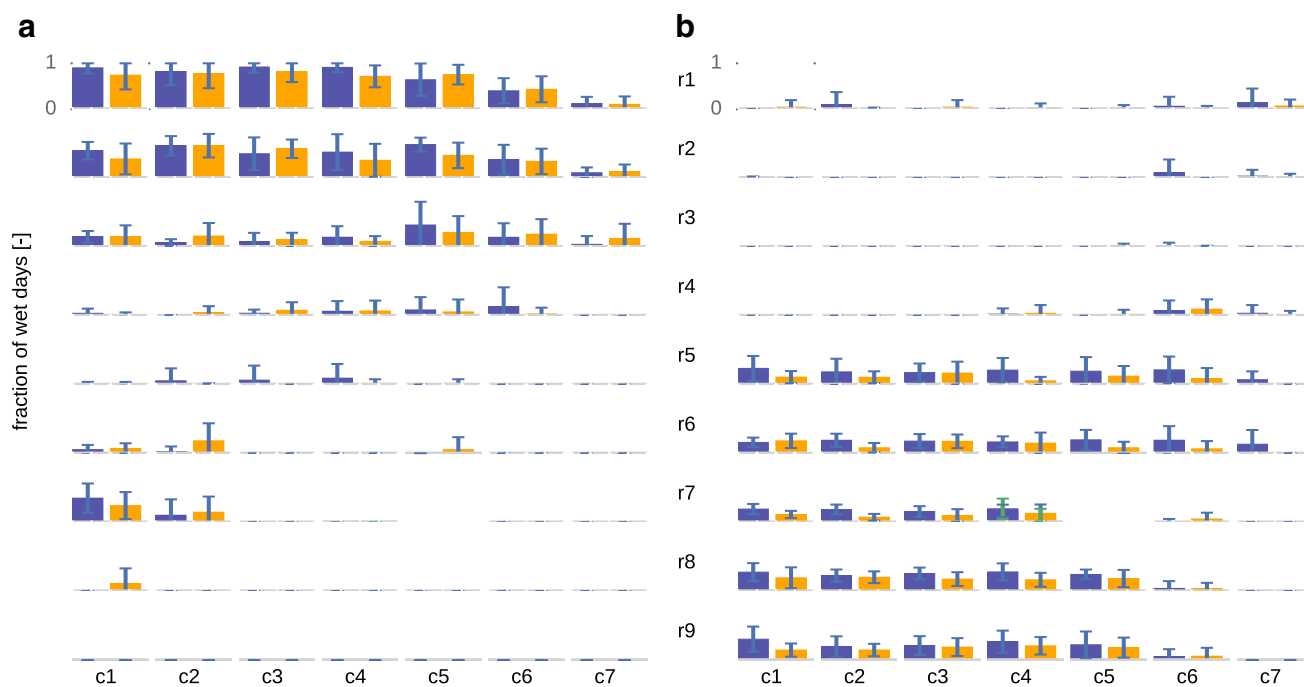


Fig. 6 Proportion of rain days per days characterized by synoptic conditions schematized by SOM shown in Fig. 2, for **a** Cape Town (Maitland) and **b** Johannesburg (Pretoria WO). Bars are mean value

over the 10 wettest (in blue) and 10 driest (in orange) JJA or DJF seasons in the 1979–2015 period. Error bars denote ± 1 standard deviation range

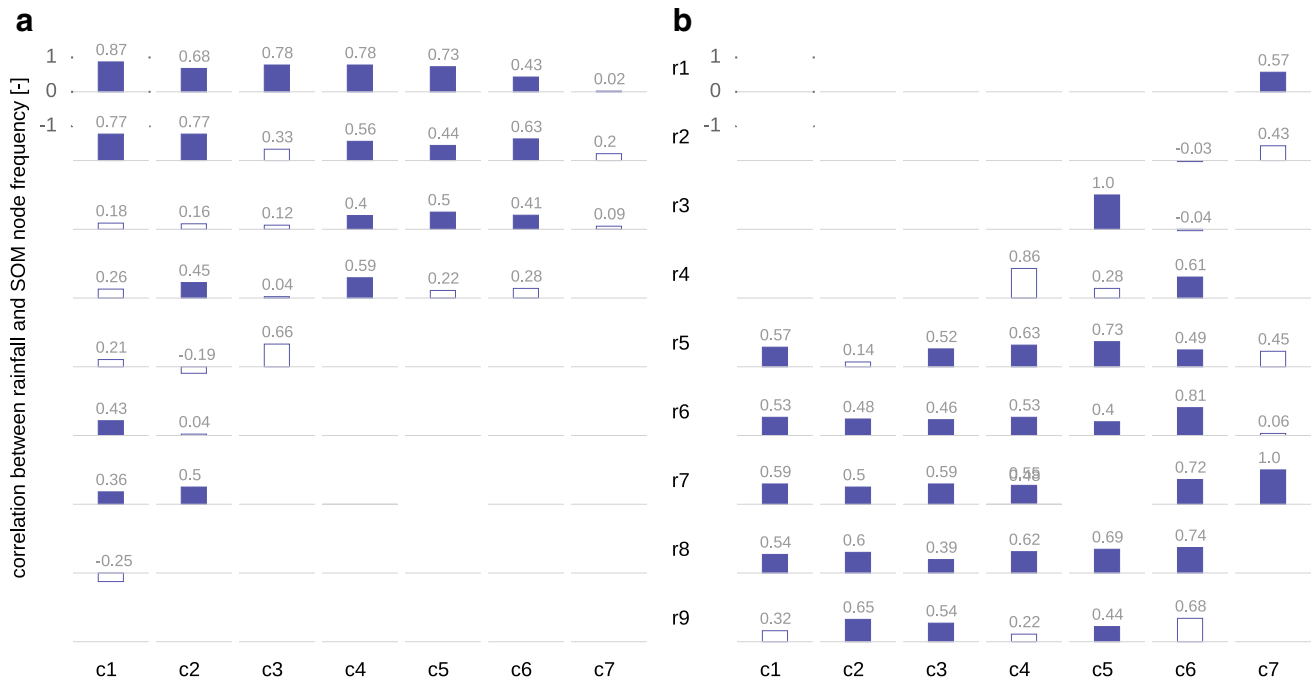


Fig. 7 Correlation between time series of seasonal synoptic state frequency and that of accumulated seasonal rainfall for synoptic states schematized by SOM shown in Fig. 2, for **a** Cape Town (during JJA)

and **b** Johannesburg (during DJF). Correlations significant at $p=0.05$ level are *solid bars*, correlations not significant at $p=0.05$ are *empty bars*

2.3 Relationship between per-node seasonal rainfall and SOM node frequency

For both domains, the strength of the relationship between the frequency of synoptic states (nodes) and total seasonal rainfall, are illustrated by correlations in Fig. 7. As one may expect, correlations are positive because a higher frequency of a rain-bearing synoptic state obviously results in a higher total seasonal rainfall from that state. The differences in magnitude of correlations between nodes indicate, however, that for some synoptic states the synoptic state–rainfall relationship is strong (i.e. when a given state occurs, a rain event of a particular magnitude is likely), while for others it is weak (i.e. when a given state occurs, it may or may not rain). This is likely due to a combination of two factors: differences between synoptic states in the magnitude of the random component in rainfall response to a given synoptic forcing, or simply the inability of a particular SOM to adequately capture the systematic component of rainfall forcing. Our analysis does not allow for an unequivocal attribution of causes of this effect.

2.4 Additional explanatory variables of interannual rainfall variability

Figure 8 illustrates the total variance in the time series of total seasonal rainfall explained by the SOM node

frequency-only model, and in combination with an additional 12 individual explanatory variables for each of the analysed large domains. Each result is obtained from 20 different realizations of the SOM training procedure and the results clearly show that SOM node frequency alone (bars corresponding to “*pr~freq*”) is in general a relatively inconsistent predictor of seasonal rainfall total. For Cape Town, the frequency of synoptic states from a SOM based on the large domain (*wcp*) explains a considerable portion (40–60%) of rainfall variance, while for Johannesburg and the *saf* domain, its explanatory power is poor explaining only 1–40% of rainfall variance.

Including local explanatory variables through piecewise linear regression increases the explained variance, depending on location and variable, so the total explained variance may reach 50–75%. The additional explanatory variables with greatest explanatory power are related to atmospheric pressure over Cape Town (*z700*, *z850* and *msl*), and moisture availability (*t850*, *q850* and *div700*) over Johannesburg. This is consistent with the understanding of the synoptic drivers of rainfall over Cape Town, which is dominated by mid-latitude low pressure systems, with the depth of these systems being the primary determinant of rainfall event intensity and duration. The importance of the pressure-related variables explains why the large-scale *qtuv* SOM does not adequately differentiate between conditions where the depths of these low pressure systems is

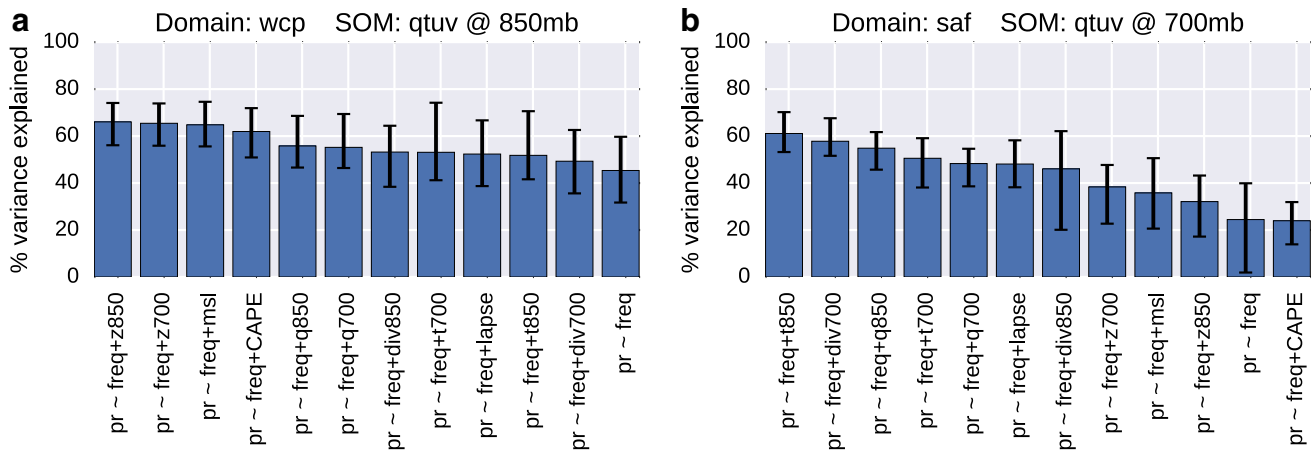


Fig. 8 Variance of total seasonal rainfall for **a** Cape Town (during JJA) and **b** Johannesburg (during DJF) explained by piecewise (node-by-node) linear regression models including *qtuv* SOM (*wcp* and *saf*

domains respectively) node frequency and an additional synoptic variable. Error bars denote a minimum–maximum range obtained from 20 different SOM training realizations

different, thus being less useful for predicting rainfall over Cape Town.

In Johannesburg, the majority of rainfall is delivered in the form of thermal-convectively driven thunderstorms. It is somewhat surprising, therefore, that a direct expression of convection potential, *CAPE*, turns out to be the least significant factor affecting interannual variability. The importance of temperature, humidity and moisture divergence suggests, therefore, that the drivers of interannual variability are not so much related to instability and convection, but rather to the levels of available moisture. In spite of the fact that the SOM includes two of the critical variables (*q* and *t*), the large domain *qtuv* SOM does not clearly differentiate between different moisture availability directly over

Johannesburg. This is likely because spatial differences in these variables over the wider region dominate the SOM training, subduing the variance arising through temporal variability over the Johannesburg region.

Figure 9 (for Cape Town) and 10 (for Johannesburg) show the comparison of the observed time series of total seasonal rainfall, with the time series predicted using piecewise linear regression and SOM node frequencies only, as well as that using the “best” additional explanatory variable. In both cases, the relationship between total seasonal rainfall and the regression residual (middle panel in Figs. 9, 10) and the comparison of regression residuals for the 10 wettest and 10 driest years indicate that the regression model systematically underestimates rainfall values in wet

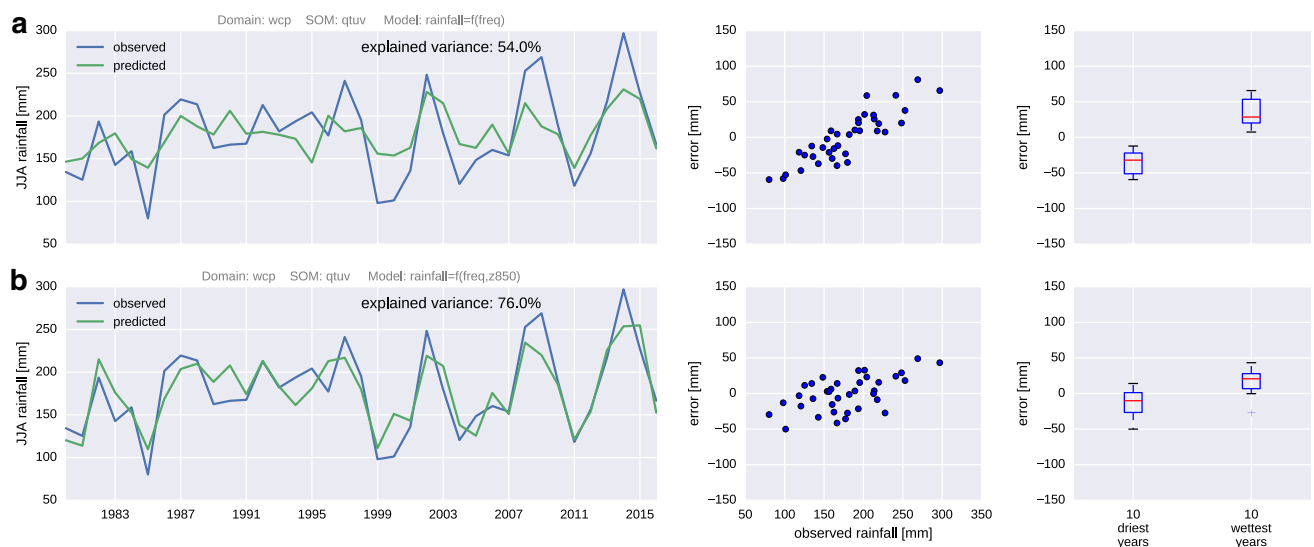


Fig. 9 Piecewise regression model for Cape Town JJA rainfall based on **a** SOM node frequency only and **b** node frequency and z850

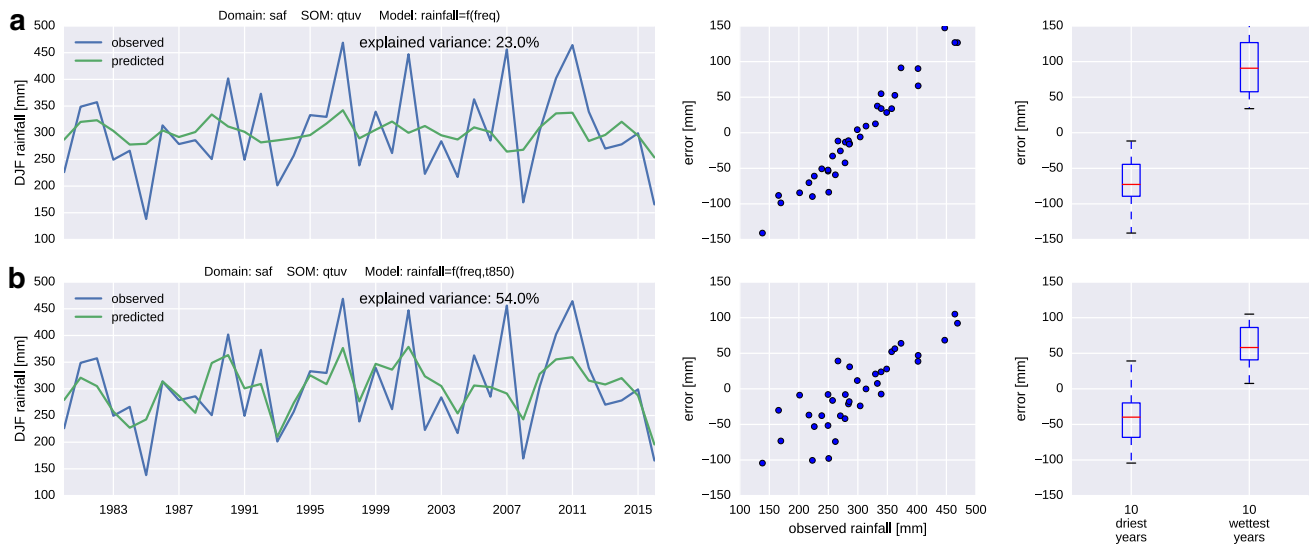


Fig. 10 Piecewise regression model for Johannesburg DJF rainfall based on **a** SOM node frequency only and **b** node frequency and t850

years, and overestimates them in dry years. This suggests the presence of a systematic component of variance in total seasonal rainfall that is not captured by the regression. It is evident that the additional explanatory variable reduces that systematic component, but does not remove it entirely.

2.5 Role of modalities of SOM procedure

Here we systematically compare the explanatory power of SOM node frequency for SOM trained with different sets of synoptic variables and domain sizes, and illustrate gains in the explanatory power provided by inclusion of additional variables into the rainfall–node frequency regression.

In Cape Town, the difference in the explanatory power of node frequency in the *wcp* (large domain) *qtuv* SOM and that in the *cpt* (small domain) *qtuv* SOM is small for *q*, *t*, *u* and *v* at 850 hPa (~45% for both domains), but the *cpt* SOM appears to be better than the *wcp* one when these variables are at 700 hPa (45 vs 30%, Fig. 11a).

For Johannesburg, node frequency of the SOM based on *q*, *t*, *u*, *v* at 700 hPa is a relatively poor predictor of total seasonal rainfall for both the *jhb* (small domain) and *saf* (large domain) SOM (~15–25% variance in seasonal rainfall explained, Fig. 11a). However, the node frequency in both *jhb* and *saf* SOM with variables at 850 hPa explain a high proportion of variance ~45% (Fig. 11a).

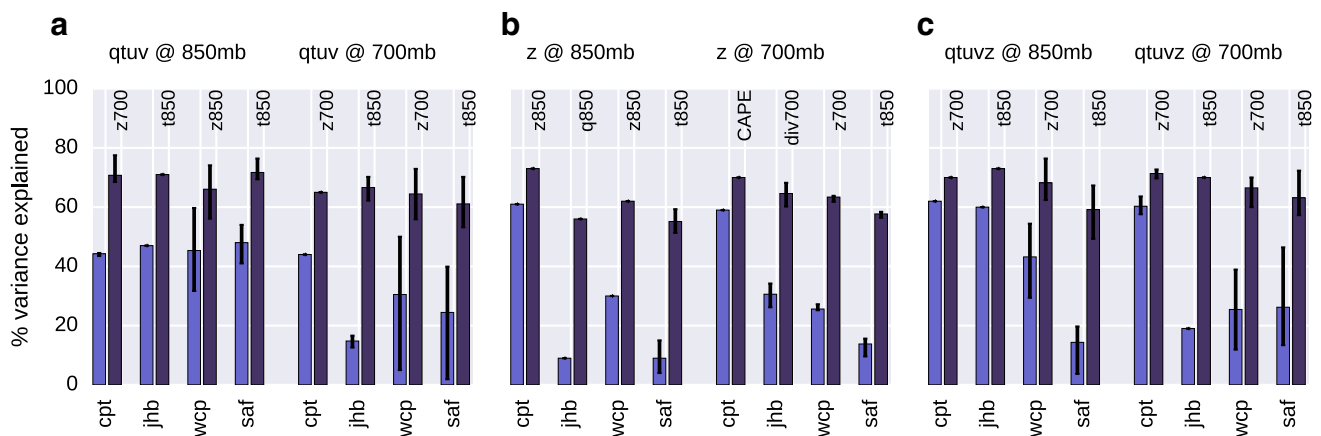


Fig. 11 Comparison of variance in total seasonal rainfall explained by piecewise linear models with node frequency (light blue bars) and by a two-variable model with frequency and an additional synoptic variable (dark blue bars) in **a** *qtuv* SOM, **b** *z* SOM and **c** *qtuvz* SOM. The two-variable model shown is one that gives the highest explained

variance—the included variable is indicated above the bar. Error bars denote min–max range of values obtained from 20 different SOM training realizations. Lack of error bars indicates that the different training realizations converged on an identical SOM

For the SOM based on geopotential height (z) only, there is a considerable increase in the explanatory power of node frequency, compared to that in q_{tuv} SOM, only for cpt SOM for Cape Town (compare light blue bars in Fig. 11b, a, ~60% for z SOM vs. ~45% for q_{tuv} SOM). For other domains, z SOM performs worse than the q_{tuv} one, explaining as low as 8% of rainfall variance. A similar result is seen for q_{tuvz} SOM (compare Fig. 11c with 11a). The high rainfall variance explained by z SOM node frequency for Cape Town is not unexpected. The earlier analyses revealed that the q_{tuv} SOM does not capture the variability in the depth of low pressure systems, and this is one of the important determinants of interannual variability in rainfall in the Cape Town region. The high level of variance in rainfall explained by z -based SOM node frequency suggests that z might indeed be the most important determinant of rainfall variability. Interestingly, the wcp (large domain) z SOM does not show a similar effect, i.e. the explanatory power does not increase when z is used in the SOM training. It is probably because the large domain SOM fails to resolve the local differences in z over the Cape Town region.

There is an increase in explanatory power of the q_{tuvz} SOM node frequency compared to that of q_{tuv} SOM for the small domain SOM, both for Cape Town and Johannesburg (Fig. 11c), but a reduction of its explanatory power in the large domain SOM at both locations. The latter is somewhat paradoxical, as all the variables from the q_{tuv} SOM are also present in the q_{tuvz} SOM. It seems, therefore, that the additional information brought into the SOM by the z variable “dilutes” the SOM’s ability to capture conditions in the atmosphere that are relevant for local rainfall.

Regardless of the SOM variables and size of the domain used over Cape Town, the two-variable piecewise regression models with z_{700} , z_{850} and $CAPE$ as an additional explanatory variable have considerably better predictive power (65–75% of variance explained) than the frequency-only models (20–60% of variance explained, Fig. 11). For Johannesburg the key variables are q_{850} , t_{850} and div_{700} , and, similarly to Cape Town, the frequency–explanatory variable models increase the predictive power to the 60–75% level, with no clear differences between different sized domains (Fig. 11). It seems, however, that for z -based SOM, adding an additional explanatory variable increases explained variance only to the 55% level.

3 Discussion and conclusions

SOM is a clustering algorithm with growing use in synoptic and statistical climatology and statistical downscaling (Sheridan and Lee 2011). Perhaps the greatest advantage of SOM is the ability to take advantage of the built-in

topological constraints (similar nodes are “close” to each other in data space) to visualize a continuum of atmospheric states in a visual two dimensional “SOM map”. This allows for relatively easy interpretation of multiple statistics and variables related to synoptic states, which are potentially more challenging with alternative methods such as “traditional”, non-topological clustering (Ward’s, K-Means, etc.). Additionally, when local surface responses (rainfall, temperature, etc.) are matched with the underlying synoptic climatology, SOM provides a potentially powerful framework for exploring these response–driver relationships (Engelbrecht and Landman 2016; Lennard and Hegerl 2014). However, the degree to which this is possible and justified is seldom quantified.

The results produced and analysed here show clearly that the fraction of variance of a common surface response parameter, such as seasonal rainfall, explained by the SOM node frequencies can be remarkably low (as low as 8% of variance explained). Even in cases where the SOM node frequencies explain relatively high portions of variance, that portion remains below 60%.

The residual, unexplained variance, also has a strong systematic component, significant fractions of which can readily be explained by one or two local circulation variables such as 850 hPa geopotential height through a simple linear model. This indicates that the SOM by itself is failing to capture significant fractions of systematic variance underlying local rainfall processes, even when small domains are utilised and the most useful explanatory variables are included in the SOM. Moreover, we show that even with two-variable models (i.e. SOM node frequency and a local circulation variable), the remaining variance, in the two cases studied here, may reach 30% and it may still contain some systematic component.

The two dimensional topological constraint of the SOM method means that in the local phase space of a particular cluster/node the intra-node variance is largely constrained to two axis and the alignment of these axis is strongly dominated by the topological constraint (i.e. the archetype of the topologically neighbouring nodes). Similar to other clustering approaches, the SOM is merely attempting to identify a range of archetype states rather than optimize explanatory power with respect to some other unconsidered variable or statistic such as seasonal rainfall totals. However, the SOM is significantly more constrained by the topological constraint than most other clustering methods and it seems possible, though it would require further analysis to confirm this, that alternative clustering methods would yield higher explained variance under a similar analysis.

While it is possible to build more complex linear (or non-linear) models involving multiple explanatory variables, we refrain from doing so in this paper for two reasons. Firstly, the purpose of the paper is to explore the

consequences of choosing different SOM configurations to elucidate the relationship between SOM synoptic classification and rainfall (and to identify the main factors that the SOM-based classification may be missing in this regard). This is to provide a deeper understanding of the interpretation of SOM synoptic classifications used to assess climate model simulations and their biases. In this context, the determination of the exact multivariate quantitative relationships which maximise explained variance is of less value.

Secondly, multiple regression modelling, although beneficial in terms of maximizing the explanatory power of the predictor–predictand relationship, may lead to its over-optimization. As such these relationships may not be stationary outside the particular conditions for which the model was calibrated, particularly under anthropogenic climate change.

Based on our results, it seems clear that the quantification of the percentage of variance explained by node frequencies should be a foundational step in any SOM-based, or for that matter any clustering method, which seeks to interpret the frequency of synoptic states in terms of drivers of local climate.

Acknowledgements This work was conducted under Future Resilience for African Cities and Lands (FRACTAL) project, which is part of the Future Climate for Africa (FCFA) program funded by UK's Department of International Development and National Environmental Research Council.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Brown JR, Jakob C, Haynes JM et al (2010) An evaluation of rainfall frequency and intensity over the Australian region in a global climate model. *J Clim* 23:6504–6525. doi:[10.1175/2010JCLI3571.1](https://doi.org/10.1175/2010JCLI3571.1)
- Cassano EN, Cassano JJ (2010) Synoptic forcing of precipitation in the Mackenzie and Yukon River basins. *Int J Climatol* 30:658–674. doi:[10.1002/joc.1926](https://doi.org/10.1002/joc.1926)
- Cavazos T (2000) Using Self-Organizing Maps to Investigate Extreme Climate Events : An Application to Wintertime Precipitation in the Balkans. *J Clim* 13:1718–1732
- Cavazos T, Hewitson BC (2005) Performance of NCEP—NCAR reanalysis variables in statistical downscaling of daily precipitation. *Clim Res* 28:95–107
- Dee DP, Uppala SM, Simmons AJ et al (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q J R Meteorol Soc* 137:553–597. doi:[10.1002/qj.828](https://doi.org/10.1002/qj.828)
- Engelbrecht CJ, Landman WA (2016) Interannual variability of seasonal rainfall over the Cape south coast of South Africa and synoptic type association. *Clim Dyn* 47:295–313. doi:[10.1007/s00382-015-2836-2](https://doi.org/10.1007/s00382-015-2836-2)
- Engelbrecht CJ, Landman WA, Engelbrecht FA, Malherbe J (2015) A synoptic decomposition of rainfall over the Cape south coast of South Africa. *Clim Dyn* 44:2589–2607. doi:[10.1007/s00382-014-2230-5](https://doi.org/10.1007/s00382-014-2230-5)
- Favre A, Hewitson B, Lennard C, et al (2013) Cut-off Lows in the South Africa region and their contribution to precipitation. *Clim Dyn* 41:2331–2351. doi:[10.1007/s00382-012-1579-6](https://doi.org/10.1007/s00382-012-1579-6)
- Finnis J, Cassano JJ, Holland M et al (2009) Synoptically forced hydroclimatology of major Arctic watersheds in general circulation models; Part 2: Eurasian watersheds. *Int J Climatol* 29:1244–1261. doi:[10.1002/joc.1769](https://doi.org/10.1002/joc.1769)
- Harrison MSJ (1984) A generalized classification of South African summer rain-bearing synoptic systems. *J Climatol* 4:547–560. doi:[10.1002/joc.3370040510](https://doi.org/10.1002/joc.3370040510)
- Hart NCG, Reason CJC, Fauchereau N (2013) Cloud bands over southern Africa: seasonality, contribution to rainfall variability and modulation by the MJO. *Clim Dyn* 41:1199–1212. doi:[10.1007/s00382-012-1589-4](https://doi.org/10.1007/s00382-012-1589-4)
- Hewitson BC, Crane RG (2002) Self-organizing maps: applications to synoptic climatology. *Clim Res* 22:13–26. doi:[10.3354/cr022013](https://doi.org/10.3354/cr022013)
- Hewitson BC, Crane RG (2006) Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. *Int J Climatol* 26:1315–1337. doi:[10.1002/joc.1314](https://doi.org/10.1002/joc.1314)
- Higgins ME, Cassano JJ (2010) Response of Arctic 1000 hPa circulation to changes in horizontal resolution and sea ice forcing in the Community Atmospheric Model. *J Geophys Res* 115:D17114. doi:[10.1029/2009JD013440](https://doi.org/10.1029/2009JD013440)
- Kohonen T (2001) *Self-Organizing Maps*. Springer Berlin Heidelberg, Berlin
- Lennard C, Hegerl G (2014) Relating changes in synoptic circulation to the surface rainfall response using self-organising maps. *Clim Dyn* 44:861–879. doi:[10.1007/s00382-014-2169-6](https://doi.org/10.1007/s00382-014-2169-6)
- Ohba M, Kadokura S, Nohara D, Toyoda Y (2016) Rainfall downscaling of weekly ensemble forecasts using self-organising maps. *Tellus, Ser A Dyn Meteorol Oceanogr* 68:29293. doi:[10.3402/tellusa.v68.29293](https://doi.org/10.3402/tellusa.v68.29293)
- Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal* 49:974–997. doi:[10.1016/j.csda.2004.06.015](https://doi.org/10.1016/j.csda.2004.06.015)
- Sammon JW (1969) A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans Comput C* 18:401–409. doi:[10.1109/T-C.1969.222678](https://doi.org/10.1109/T-C.1969.222678)
- Schuenemann KC, Cassano JJ, Finnis J (2008) Synoptic forcing of precipitation over Greenland climatology for 1961–99. *J Hydrometeorol* 10:60–78. doi:[10.1175/2008JHM1014.1](https://doi.org/10.1175/2008JHM1014.1)
- Sheridan SC, Lee CC (2011) The self-organizing map in synoptic climatological research. *Prog Phys Geogr* 35:109–119. doi:[10.1177/0309133310397582](https://doi.org/10.1177/0309133310397582)
- Tyson PD, Preston-White RA (2000) *The weather and climate of southern Africa*. Oxford University Press, Oxford
- Verdon-Kidd DC, Kiem AS (2009) On the relationship between large-scale climate modes and regional synoptic patterns that drive Victorian rainfall. *Hydrol Earth Syst Sci* 13:467–479. doi:[10.5194/hessd-5-2791-2008](https://doi.org/10.5194/hessd-5-2791-2008)
- Yin C, Li Y, Ye W et al (2011) Statistical downscaling of regional daily precipitation over southeast Australia based on self-organizing maps. *Theor Appl Climatol* 105:11–26. doi:[10.1007/s00704-010-0371-y](https://doi.org/10.1007/s00704-010-0371-y)